# EDirect: Command Line Access to NCBI Databases

Searching, retrieving, and parsing data from NCBI databases through the Unix command line

**https://www.ncbi.nlm.nih.gov/books/NBK179288/**

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

## Introduction

Entrez Direct (EDirect) provides access to the NCBI's suite of interconnected databases (biomedical and life science literature, nucleotide and protein sequence, molecular structure, gene, genome assembly, gene expression, clinical variation, etc.) from a Unix terminal window. Search terms are given in command-line arguments. Individual operations are connected with Unix pipes to allow construction of multi-step queries. Selected records can then be retrieved in a variety of formats.

EDirect also includes an argument-driven function that simplifies the extraction of data from document summaries or other results that are in structured XML format. This can eliminate the need for writing custom software to answer ad hoc questions. Queries can move seamlessly between EDirect commands and Unix utilities or scripts to perform actions that cannot be accomplished entirely within Entrez.

## Programmatic Access

Several underlying network services provide access to different facets of Entrez. These include searching by indexed terms, looking up precomputed neighbors or links, filtering results by date or category, and downloading record summaries or reports. The same functionalities are available on the web or when using programmatic methods.

EDirect navigation programs (**esearch**, **elink**, **efilter**, and **efetch**) communicate by means of a small structured message, which can be passed invisibly between operations with a Unix pipe. The message includes the current database, so it does not need to be given as an argument after the first step.

All EDirect commands are designed to work on large sets of data. Intermediate results are stored on the Entrez history server. For best performance, obtain an API Key from NCBI (see reference at the end), and place the following line in your .bash_profile file:

```
export NCBI_API_KEY=user_api_key_goes_here
```

Each program also has a **-help** command that prints detailed information about available arguments.

## Navigation Functions

Esearch performs a new Entrez search using terms in indexed fields. It requires a **-db** argument for the database name and uses **-query** to obtain the search terms. For PubMed, without field qualifiers, the server uses automatic term mapping to compose a search strategy by translating the supplied query:

```
esearch -db pubmed -query "selective serotonin reuptake inhibitor"
```

Search terms can be also qualified with bracketed field names:

```
esearch -db nucleotide -query "insulin [PROT] AND rodents [ORGN]"
```

Elink looks up precomputed neighbors within a database, or finds associated records in other databases:

```
elink -related
elink -target gene
```

Efilter limits the results of a previous query, with shortcuts that can also be used in esearch:

```
efilter -molecule genomic -location chloroplast -country sweden
```

Efetch downloads selected records or reports in a designated format:

```
efetch -format abstract
```

## Entrez Exploration

Individual query commands are connected by a Unix vertical bar pipe symbol:

```
esearch -db pubmed -query "tn3 transposition immunity" | efetch -format medline
```

PubMed related articles are calculated by a statistical algorithm using the title, abstract, and medical subject headings (MeSH terms). These connections between papers can be used for knowledge discovery.

Lycopene cyclase converts lycopene to β-carotene, the immediate precursor of vitamin A. An initial search on the enzyme results in 232 articles. Looking up precomputed neighbors returns 14,387 PubMed papers, some of which might be expected to discuss adjacent steps in the biosynthetic pathway:

```
esearch -db pubmed -query "lycopene cyclase" | \
elink -related | \
elink -target protein | \
efilter -organism mouse | \
efetch -format fasta
```

Linking to the protein database finds 251,887 sequence records, each of which has standardized organism information from the NCBI taxonomy. Limiting to proteins in mice returns 39 records. (Animals do not encode the genes involved in carotene biosynthesis, except in aphids and their ilk, apparently obtained y horizontal gene transfer from fungi.) Records are then retrieved in FASTA format:

```
...
>NP_067461.2 beta,beta-carotene 15,15'-dioxygenase isoform 1 [Mus musculus]
MEIIFGQNKKEQLEPVQAKVTGSIPAWLQGTLLRNGPGMHTVGESKYNHWFDGLALLHSFSIRDGEVFYR
SKYLQSDTYIANIEANRIVVSEFGTMAYPDPCKNIFSKAFSYLSHTIPDFTDNCLINIMKCGEDFYATTE
TNYIRKIDPQTLETLEKVDYRKYVAVNLATSHPHYDEAGNVLNMGTSVVDKGRTKYVIFKIPATVPDSKK
KGKSPVKHAEVFCSISSRSLLSPSYYHSFGVTENYVVFLEQPFKLDILKMATAYMRGVSWASCMSFDRED
KTYIHIIDQRTRKPVPTKFYTDPMVVFHHVNAYEEDGCVLFDVIAYEDSSLYQLFYLANLNKDFEEKSRL
TSVPTLRRFAVPLHVDKDAEVGSNLVKVSSTTATALKEKDGHVYCQPEVLYEGLELPRINYAYNGKPYRY
IFAAEVQWSPVPTKILKYDILTKSSLKWSEESCWPAEPLFVPTPGAKDEDDGVILSAIVSTDPQKLPFLL
ILDAKSFTELARASVDADMHLDLHGLFIPDADWNAVKQTPAETQEVENSDHPTDPTAPELSHSENDFTAG
HGGSSL
...
```

As anticipated, the results include the enzyme that splits β-carotene into two molecules of retinal.

## Structured Data Extraction

The **xtract** program uses command-line arguments to direct the conversion of XML data into a tab-delimited table. The **-pattern** argument divides the results into rows, while placement of data into columns is controlled by **-element**.

Formatting arguments allow extensive customization of the output. The line break between -pattern objects can be changed with **-ret**, and the tab character between -element fields can be replaced by **-tab**. The **-sep** argument is used to distinguish multiple elements of the same type, and controls their separation independently of the -tab argument. The -sep value also applies to unrelated -element arguments that are grouped with commas. The following query:

```
efetch -db pubmed -id 6271474,1413997,16589597 -format docsum | \
xtract -pattern DocumentSummary -sep "|" -element Id PubDate Name
```

returns a table with individual author names separated by vertical bars:

```
6271474     1981        Casadaban MJ|Chou J|Lemaux P|Tu CP|Cohen SN
1413997     1992 Oct    Mortimer RK|Contopoulou CR|King JS
16589597    1954 Dec    Garber ED
```

Selection arguments are specialized derivatives of -element. Among these are positional commands (**-first** and **-last**) and numeric processing operations (including **-num**, **-len**, **-sum**, **-min**, **-max**, and **-avg**). There are also functions that perform sequence coordinate conversion (**-0-based**, **-1-based**, and **-ucsc-based**).

## Nested Exploration

Exploration arguments (**-pattern**, **-group**, **-block**, and **-subset**) limit data extraction to specified regions of the XML, visiting all relevant objects one at a time. This design allows nested exploration of complex, hierarchical data to be controlled by a linear chain of command-line argument statements.

PubmedArticle XML contains the MeSH terms applied to a publication. Each MeSH term can have its own unique set of qualifiers. A single level of nested exploration within the current pattern:

```
esearch -db gene -query "beta-carotene oxygenase 1" -organism human | \
elink -target pubmed | efilter -released last_year | efetch -format xml | \
xtract -pattern PubmedArticle -element MedlineCitation/PMID \
  -block MeshHeading \
    -pfc "\n" -sep "/" -element DescriptorName,QualifierName
```

retains the proper association of subheadings for each MeSH term:

```
30396924
Age Factors
Animals
Cell Cycle Proteins/deficiency/genetics/metabolism
...
```

A second level (-subset) would be needed to print major topic attributes next to their parent subheadings.

## Conditional Execution

Conditional processing arguments (**-if** and **-unless**) restrict exploration by object name and value. These may be used in conjunction with string or numeric constraints:

```
esearch -db pubmed -query "Casadaban MJ [AUTH]" | \
efetch -format xml | \
xtract -pattern PubmedArticle -if "#Author" -lt 6 \
  -block Author -if LastName -is-not Casadaban \
    -sep ", " -tab "\n" -element LastName,Initials | \
sort-uniq-count-rank
```

to select papers with fewer than 6 authors and print a table of the most frequent coauthors:

```
11    Chou, J
8     Cohen, SN
7     Groisman, EA
...
```

## Saving Data in Variables

A value can be recorded in a variable and used wherever needed. Variables are created by a hyphen followed by a name consisting of a string of capital letters or digits (e.g., **-PMID**). Values are retrieved by placing an ampersand (&) before the variable name (e.g., "**&PMID**") in an -element statement:

```
efetch -db pubmed -id 3201829,6301692,781293 -format xml | \
xtract -pattern PubmedArticle -PMID MedlineCitation/PMID \
  -block Author -element "&PMID" \
    -sep " " -tab "\n" -element Initials,LastName
```

producing a list of authors, with the PubMed Identifier (PMID) in the first column of each row:

```
3201829    JR Johnston
3201829    CR Contopoulou
3201829    RK Mortimer
6301692    MA Krasnow
6301692    NR Cozzarelli
781293     MJ Casadaban
```

The variable can be used even though the original object is no longer visible inside the -block section.

## Sequence Qualifiers

The NCBI represents sequence records in a data model based on the central dogma of molecular biology. A sequence can have multiple features, which carry information about the biology of a given region, including the transformations involved in gene expression. A feature can have multiple qualifiers, which store specific details about that feature (e.g., name of the gene, genetic code used for translation).

The data hierarchy is easily explored using a **-pattern** {sequence} **-group** {feature} **-block** {qualifier} construct. As a convenience, an **-insd** helper function is provided for generating the appropriate nested extraction commands from feature and qualifier names on the command line. For example, processing the results of a search on cone snail venom:

```
esearch -db protein -query "conotoxin" -feature mat_peptide | \
efetch -format gpc | \
xtract -insd complete mat_peptide "%peptide" product peptide | \
grep -i conotoxin | sort -t $'\t' -u -k 2,2n
```

returns the accession, length, name, and sequence for a sample of neurotoxic peptides:

```
ADB43131.1    15    conotoxin Cal 1b      LCCKRHHGCHPCGRT
ADB43128.1    16    conotoxin Cal 5.1     DPAPCCQHPIETCCRR
AIC77105.1    17    conotoxin Lt1.4       GCCSHPACDVNNPDICG
ADB43129.1    18    conotoxin Cal 5.2     MIQRSQCCAVKKNCCHVG
ADD97803.1    20    conotoxin Cal 1.2     AGCCPTIMYKTGACRTNRCR
AIC77085.1    21    conotoxin Bt14.8      NECDNCMRSFCSMIYEKCRLK
ADB43125.1    22    conotoxin Cal 14.2    GCPADCPNTCDSSNKCSPGFPG
AIC77154.1    23    conotoxin Bt14.19     VREKDCPPHPVPGMHKCVCLKTC
...
```

## Additional Information

**Full documentation**
The help manual for Entrez Programming Utilities is at:
https://www.ncbi.nlm.nih.gov/books/NBK25501/
Full EDirect help manual is at:
https://www.ncbi.nlm.nih.gov/books/NBK179288/

**Video tutorials**
NCBI video tutorials on EDirect are available in NCBI YouTube channel:
https://www.youtube.com/user/NCBINLM/search?query=EDirect
An NLM course on EDirect is linked off this page:
https://dataguide.nlm.nih.gov/classes.html

**Short handouts**
Two EDirect installation handouts are at:
https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_EDirect_Unix.pdf
https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_EDirect_Windows.pdf
A booklet with practical usage examples is at:
https://ftp.ncbi.nlm.nih.gov/pub/factsheets/Booklet_Teaching_EDirect.pdf
Information on how to obtain an API key is described in this NCBI blogpost:
https://ncbiinsights.ncbi.nlm.nih.gov/2017/11/02/new-api-keys-for-the-e-utilities/
For information on other NCBI resources, please refer to our factsheets collection and online help manuals:
http://bit.ly/ncbi_factsheets
https://www.ncbi.nlm.nih.gov/books/NBK3831/

**Email contact**
Please subscribe to the utilities announce mailing list to get informed on pending changes and updates:
https://www.ncbi.nlm.nih.gov/mailman/listinfo/utilities-announce
We look forward to hearing from our users, please send your questions and comments to:
info@ncbi.nlm.nih.gov